Formulación de Esquemas de Almacenamiento de Datos Médicos para aplicar Minería de Datos en el Diagnóstico de Enfermedades

Ana Lía Carabio[1], Elizabeth Silva Layes[1], Marcelo A. Falappa[2]

[1]Facultad de Ciencias de la Administración - Universidad Nacional de Entre Ríos
Monseñor Tavella 1424 – Concordia, Entre Ríos (3200) - Tel.: +54(0345)4231406
anacar@fcad.uner.edu.ar, elizabeth.silva@gmail.com

[2]Departamento de Ciencias e Ingeniería de la Computación - Universidad Nacional del Sur
San Andrés 800 – Campus de Palihue - Bahía Blanca (8000) - Tel.: +54(0291)4595135
mfalappa@cs.uns.edu.ar

Resumen

El sector salud administra grandes volúmenes de datos, centrándose la mayoría de las tomas de decisiones en el área clínica. Por tal motivo, contar con información útil, inmediata y efectiva es sumamente relevante en éste ámbito.
En este sentido, la minería de datos es una herramienta que permite encontrar patrones de comportamiento de utilidad para la toma de decisiones clínicas, como lo son la realización de estudios epidemiológicos, cálculo de expectativas de vida, identificación de terapias médicas satisfactorias para diferentes enfermedades, entre otros.
El objetivo del presente trabajo es construir un esquema que, a partir de la extracción de datos médicos relevantes de las Historias Clínicas Electrónicas (HCE), permita almacenarlos de manera eficiente en una base de datos NoSQL, como *HBase*, con la finalidad de aplicar técnicas de minería de datos.

Palabras clave: Historia Clínica Electrónica, Minería de Datos, Big Data, Bases de Datos NoSQL, HBase, Hadoop.

Contexto

Este trabajo se desarrolla dentro del Proyecto de Investigación y Desarrollo PID 7042 “Estudio Comparativo y Análisis de Rendimiento de los Lenguajes de Manipulación de Datos en Bases de Datos Orientadas a Objetos y Bases de Datos Objeto-Relacionales”[1], cuyo período de ejecución será desde noviembre de 2014 a noviembre de 2017, en el marco de un Acuerdo de Colaboración Académico-Científico entre la Facultad de Ciencias de la Administración de la Universidad Nacional de Entre Ríos (UNER) y el Instituto de Ciencias e Ingeniería de la Computación (ICIC) del Departamento de Ciencias e Ingeniería de la Computación (DCIC) de la Universidad Nacional del Sur (UNS).
Uno de los objetivos del proyecto apunta a establecer comparaciones en el rendimiento de sistemas desarrollados en

lenguajes orientados a objetos que interactúan con diversos modelos de bases de datos.

Además, este proyecto prioriza la formación de recursos humanos para investigación en la Facultad de Ciencias de la Administración de la UNER, especializados en la línea de investigación denominada "Ingeniería de Software y Lenguajes de Programación" establecida por Res. 25/11 del Consejo Directivo.

Introducción

En la actualidad, el procesamiento de grandes volúmenes de datos (*Big Data*) para la toma de decisiones ha dejado de ser privativo de organizaciones comerciales y de negocios, y se ha inmiscuido en otros ámbitos, con actividades variadas y diversos intereses. Entre ellos, el sector salud es uno de los sectores que más se ha visto beneficiado con la utilización de herramientas de análisis de datos.

La vasta acumulación de datos clínicos existentes en los *Electronic Health Records* (EHR's) [1] (diagnósticos, tratamientos indicados, paraclínicas, medicamentos suministrados, procedimientos realizados, etc.) presentes en la mayoría de las instituciones sanitarias, brinda una oportunidad inmejorable para: la realización de estudios epidemiológicos, el cálculo de expectativas de vida, la identificación de terapias médicas satisfactorias para diferentes enfermedades, etc [2].

El sector sanitario, en su totalidad, es uno de los que administra los mayores volúmenes de datos, centrándose la mayoría de las tomas de decisiones en el área clínica. Sin duda, los Sistemas de Soporte a Decisiones Clínicas (CDSS), además de apoyar al médico en la toma de decisiones vinculadas a diagnósticos, protocolos clínicos que se deben activar ante un diagnóstico, medicación y/o procedimientos a prescribirse a un paciente, también deben verse como un soporte para la prevención de enfermedades [3]. Por esta razón, contar con información útil, inmediata y efectiva es sumamente relevante en éste ámbito.

En este sentido, la minería de datos es una herramienta que permite encontrar patrones de comportamiento de utilidad para la toma de decisiones vinculadas a este último punto.

La minería de datos se relaciona de manera estrecha con la estadística, utilizando técnicas de muestreo y visualización de datos, y depuración y cálculo de indicadores, entre otros. Según Hand et al. [4], "la minería de datos es el análisis de grandes conjuntos de datos observacionales para encontrar relaciones insospechadas, y para resumir los datos en nuevas formas, comprensibles y útiles para el titular de los datos".

Entre los obstáculos que puede encontrar la aplicación de minería de datos en la medicina, podemos mencionar la voluminosidad y heterogeneidad de los datos médicos, la complejidad de su representación, la posible incompletitud de los mismos, entre otros. Esto fuerza a las instituciones sanitarias a realizar grandes inversiones en tiempo y dinero para poder procesar esta información adecuadamente [5].

En lo que se refiere al tratamiento de la

voluminosidad y heterogeneidad de los datos, el mismo se ha visto mejorado por la aparición de las bases de datos NoSQL. En particular, las del tipo orientadas a columnas (*Column-Oriented Databases*), adecuadas para aplicarlas en minería de datos y aplicaciones analíticas, por su forma de almacenamiento, la compresión eficiente de los datos y por su diseño, que permite cargar y analizar grandes volúmenes de datos [6, 7, 8, 9].

Considerando lo enunciado en [6] que, *"…las bases de datos orientadas a columnas son adecuadas para aplicaciones analíticas y de minería de datos, donde el método de almacenamiento es ideal para las operaciones comunes que se realizan en los datos…"*, se utilizará *HBase* sobre *Hadoop* para almacenar los datos relevantes que se pretenden obtener de las HCE [10, 11].

El proyecto *Apache Hadoop*, de la *Apache Software Foundation*, es uno de los enfoques existentes para el análisis de datos no estructurados. *Hadoop* es un *framework open source* que permite el procesamiento distribuido de grandes conjuntos de datos a través de *clusters* de computadoras, ofreciendo escalabilidad y confiabilidad [12].

HBase [13], que también forma parte del proyecto *Apache Hadoop*, es un sistema de gestión de bases de datos orientado a columnas que se ejecuta en la parte superior del HDFS [12], que no admite un lenguaje de consulta como SQL (*Structured Query Language*), y que se utiliza con frecuencia para analizar grandes conjuntos de datos. Es una base de datos distribuida, no relacional, *open-source* modelada a partir de *Google Big Table*[1], y que soporta scripts escritos en Java [12].

Líneas de Investigación, Desarrollo e Innovación

En la actualidad han cobrado importancia las bases de datos no puramente relacionales, caracterizadas, principalmente, por su almacenamiento distribuido y su fácil escalabilidad. En esta línea, se buscará analizar el comportamiento de una base de datos del tipo NoSQL como lo es *HBase*, al aplicar herramientas de minería de datos, con la finalidad de evaluar su rendimiento ante la necesidad de analizar grandes volúmenes de datos.

Entendiendo que uno de los campos de aplicación fértiles de la minería de datos es el campo de la bioingeniería, nos proponemos integrar el conocimiento que formula la minería de datos a la HCE como apoyo en la toma de decisiones clínicas, implementando un esquema que permita obtener y generar un repositorio con datos relevantes a fin de agilizar la obtención de resultados.

Resultados y Objetivos

Dada la importancia que en la actualidad está presentando el manejo de grandes volúmenes de datos, y la importancia que están adquiriendo las bases de datos NoSQL, se hace necesario integrar la

[1] *Bigtable* es un sistema de almacenamiento distribuido para gestionar datos estructurados, diseñado para escalar a un tamaño muy grande utilizado por *Google* [14].

utilización de este tipo de BD para manejar la variedad y complejidad de los datos médicos al aplicar herramientas de análisis. Para ello se prevé:

- Instalar y configurar *Hadoop* y *HBase*. En una primera etapa, se instalará y configurará *Hadoop* de una forma pseudo-distribuida. Luego se instalará y confirgurará *HBase* sobre la misma instalación.
- Obtener datos relevantes desde una base de datos relacional que contiene las HCE. Para ello se diseñará y desarrollará una interface en lenguaje *Java* [15] que permita extraer los datos relevantes vinculados al área de interés de análisis y volcarlos a la base de datos generada en *HBase*.
- Evaluar el correcto funcionamiento del proceso desarrollado. Para realizar las pruebas se trabajará con los datos obtenidos y referidos en [9], a fin de probar la correctitud del mismo.
- Aplicar técnicas de minería de datos sobre la base de datos NoSQL generada, con el objetivo de analizar el tema de interés clínico. Se utilizará el software *WEKA*[2] y los resultados se cotejarán con los resultados obtenidos en [9], para verificar la exactitud de los datos manipulados.

Finalmente, se pretende incorporar este esquema a un componente desarrollado en *Java* que puede prestar servicios e integrarse a una HCE para brindar apoyo en el diagnóstico médico.

Formación de Recursos Humanos

Como parte del actual proyecto de investigación se espera que uno de los docentes investigadores, y que es autor de este artículo, complete su Tesis de Magister en Redes de Datos en la Facultad de Informática de la Universidad Nacional de La Plata. También, se espera que otro de los autores de este proyecto complete su Tesis de Magister en Sistemas de Información en la Facultad de Ciencias de la Administración de la Universidad Nacional de Entre Ríos. Finalmente, se buscará formar nuevas sublíneas de investigación relacionadas a este proyecto, así como también la formación de nuevos alumnos en los posgrados dictados en el ámbito de la Universidad Nacional de Entre Ríos y de la Universidad Nacional del Sur.

Referencias

[1] Balas, E. A., Vernon, M., Magrabi, F., Gordon, L. T. & Sexton, J. (2015). *Big Data Clinical Research: Validity, Ethics, and Regulation.* **In MEDINFO 2015: EHealth-enabled Health: Proceedings of the 15th World Congress on Health and Biomedical Informatics**, Vol. 216. IOS Press, 448.

[2] Molina, J. & García, J. (2006). *Técnicas de análisis de datos: Aplicaciones prácticas utilizando Microsoft Excel y Weka*. Universidad Carlos III de Madrid España.

[2] http://www.cs.waikato.ac.nz/ml/weka/index.html

[3] Silva Layes, M. E., Falappa, M. A., & Simari, G. R. (2). *Sistemas de Soporte a las Decisiones Clínicas*. http://sedici.unlp.edu.ar/bitstream/handle/10915/19976/Documento_completo.pdf?sequence=1.

[4] Hand, D. J. & Mannila, H., Smyth, P. (2001). *Principles of Data Mining*. **MIT press** ISBN: 026208290x.

[5] Milovic, B.& Milovic, M. (2012). *Prediction and Decision Making in Health Care using Data Mining*. **International Journal of Public Health Science (IJPHS)**, Vol. 1, Nº 2, 69-78 ISSN: 2252-8806.

[6] Nayak, A., Poriya, A. & Poojary, D. (2013). *Type of NOSQL databases and its comparison with relational databases*. **International Journal of Applied Information Systems**, Vol. 5, Nº 4, 16-19.

[7] Mehta, R. G., Mistry, N. J. & Raghuvanshi, M. (2013). *Impact of Column-oriented Databases on Data Mining Algorithms*. **International Journal of Advanced Research in Computer and Communication Engineering**.

[8] Carabio, A. L. R., Benedetto, M. G. & Falappa, M. A. (2016). *Comportamiento de Bases de Datos No Relacionales en Entornos Distribuidos*. **In XVIII Workshop de Investigadores en Ciencias de la Computación, WICC'2016**.

[9] Carabio, A. L. R., Silva Layes, M. E., Frola, F., & Falappa, M. A. (2016). *Bioingeniería aplicada en el diagnóstico de enfermedades*. **In VII Congreso Argentino de Informática en Salud (CAIS 2016)-JAIIO 45**.

[10] Das, T. K., & Kumar, P. M. (2013). *Big Data Analytics: A framework for unstructured data analysis.* **International Journal of Engineering and Technology (IJET)**, 5(1), 153-156.

[11] Song, H., Li, L., & Fan, Y (2014). *Applied research on data mining platform for weather forecast based on cloud storage*. **Computer Modelling & New Technologies**, 18(12C) 1226-1230.

[12] The *Apache Hadoop Project*, http://hadoop.apache.org/ (2016).

[13] *Apache HBase Project*, http://hbase.apache.org/ (2016).

[14] Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., ... & Gruber, R. E. (2006). *Bigtable: A Distributed Storage System for Structured Data*. **To appear in OSDI,** 1.

[15] *Java Platform,* Standard Edition (Java SE) 8, http://docs.oracle.com/javase/8/index.html (2016).